Assessing Writing xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system

Elaine Lin Wang[a,*], Lindsay Clare Matsumura[b], Richard Correnti[b], Diane Litman[b], Haoran Zhang[b], Emily Howe[b], Ahmed Magooda[b], Rafael Quintana[b]

[a] *RAND Corporation, 4570 Fifth Avenue, Pittsburgh, PA, 15213, USA*
[b] *University of Pittsburgh, Learning Research and Development Center, 3939 O'Hara Street, Pittsburgh, PA, 15260, USA*

ARTICLE INFO

Keywords:
Writing
Argument writing
Evidence use
Formative feedback
Automated writing evaluation
Automated essay scoring

ABSTRACT

We investigate students' implementation of the feedback messages they received in an automated writing evaluation system (*eRevise*) that aims to improve students' use of text evidence in their writing. Seven 5th and 6th-grade teachers implemented *eRevise* (n = 143 students). Qualitative analysis of students' essays across first and second drafts suggests that the majority of students made changes to their essays that were in line with the feedback they received, though few of these changes resulted in substantive improvement in essay quality. Twenty percent of students did not attempt to implement the feedback; these students generally made small changes to wording or mechanics. In response to the feedback to add more evidence, students whose essays did not improve or showed only slight improvement frequently added in evidence that was not text based or repeated evidence already present in the first draft. When prompted to explain how the evidence they included connected to their claim, many students paraphrased the evidence, added a short conclusion, or explained generally how the evidence supports claims (not how this was instantiated in their writing). Implications for teaching argument writing and for designing AWE systems that support students to successfully revise their essays are discussed.

1. Introduction

Writing is critical to learning and academic success (Bangert-Drowns, Hurley, & Wilkinson, 2004; Graham & Perin, 2007). Students who cannot write well are disadvantaged in salary positions and often find college coursework too difficult to complete (National Commission on Writing for America's Families, Schools, & Colleges, 2004). Despite the importance of writing, teachers have historically spent less time teaching writing in comparison to other subjects (Ibid.). Many teachers report feeling underprepared to teach writing well, and rarely or only partially implement research-based practices for writing instruction (Brindle, Graham, Harris, & Hebert, 2016). Unsurprisingly, results of national assessments show that the majority of students in the United States are not proficient writers (National Center for Education Statistics, 2012).

Recently, standards have begun to emphasize text-based argument writing as especially important for college readiness (e.g., Graham, Harris, & Santangelo, 2015; National Governors Association Center for Best Practices & Council of Chief State School Officers (NGAC/CCSSO), 2010). This form of writing requires students to express higher-level thinking about texts, formulate arguments, and marshal solid evidence in support of their claims. To successfully produce these essays, students must be able to comprehend source texts, produce writing, and master argument elements (e.g., claims, reasons, and evidence). Text-based argument

* Corresponding author.
E-mail address: ewang@rand.org (E.L. Wang).

https://doi.org/10.1016/j.asw.2020.100449
Received 28 May 2019; Received in revised form 12 February 2020; Accepted 13 February 2020

writing is a relatively new addition to the elementary curricula, which traditionally has emphasized narrative and creative writing (Shanahan, 2015). When elementary students do write in response to text, it is often limited to short summaries (Matsumura, Correnti, & Wang, 2015). What little research exists about the quality of text-based argument writing in the elementary grades suggests that students frequently struggle to express higher-level thinking and use text evidence well in support of claims (O'Hallaron, 2014; Wang, Matsumura, & Correnti, 2018). While students generally are successful at stating a position and marshalling a certain number of pieces of evidence, they less consistently provide detailed evidence and explain the connection between the evidence and their position (De La Paz et al., 2012; O'Hallaron, 2014; Wang et al., 2018).

Because the ability to marshal text evidence in support of a claim is a relatively new expectation for young writers, little is known about how to teach this skill well. One highly endorsed approach to developing students' writing skills broadly, however, is to engage students in cycles of planning, drafting, revising, and editing their essays (Graham & Perin, 2007; Graham & Sandmel, 2011). Key to the success of this 'process' approach to writing instruction is the provision of formative feedback on early drafts of students' work (Graham, Hebert, & Harris, 2015). Such feedback is essential for making visible important differences between current and desired practice on substantive dimensions of students' essays (e.g., argument and evidence use), for revealing key areas of improvement to guide revision, and for providing information that teachers can use to target instruction (Black & Wiliam, 1998; Heritage, 2010; Shute, 2008).

For multiple reasons, however, students rarely receive substantive formative feedback on their writing. First, teachers can be reluctant to assign writing tasks that require students to work across drafts, as providing formative feedback is time-consuming. Also, teachers vary in their implementation of evidence-based practices for teaching writing more generally (Graham, Capizzi, Harris, Hebert, & Morphy, 2014) and can feel unsure about how to provide feedback to improve students' writing (Wang et al., 2018). Moreover, research shows that when teachers do provide feedback on drafts, their edits and comments often focus on surface-level features (i.e., grammar, spelling, pronoun referents) (Matsumura, Patthey-Chavez, Valdés, & Garnier, 2002; Olson & Raffeld, 1987; Patthey-Chavez, Matsumura, & Valdes, 2004;). Students' revisions thus typically show improvement in readability, but not in content (Matsumura et al., 2002; Patthey-Chavez et al., 2004).

One approach to potentially easing the burden on teachers and increasing students' opportunities to receive substantive formative feedback is to leverage automated writing evaluation (AWE) systems. These systems combine automated essay scoring (AES) technologies with feedback on drafts of students' essays. In the present study, we describe results from a pilot study of an AWE system, *eRevise*, developed to support students' use of text evidence in argument writing.

1.1. Automated writing evaluation (AWE) systems

Evidence to date that AWE systems improve the quality of students' writing is modest (see reviews by Stevenson & Phakiti, 2014 and Graham, Hebert et al., 2015). Historically, when studies have shown positive effects, most of the improvement is in the readability and cohesion of students' responses (Attali, 2004; Kellogg, Whiteford, & Quinlan, 2010; Roscoe, Snow, Allen, & McNamara, 2015; Shermis, Garvan, & Diao, 2008; Wilson & Andrada, 2016). Notably, while AWE systems have advanced in recent years with respect to capturing features of arguments (e.g., presence or absence of a claim) (Palermo & Thomson, 2018), designing systems that assess and improve the content of students' writing is still in an early phase of development.

One explanation for why AWE systems have shown effects on limited domains of writing is that the AES systems upon which AWE systems are built mostly produce holistic scores driven by linguistic features (e.g., word count and syntax) rather than on more subjective dimensions (e.g., students' reasoning) (Deane, 2013). While this approach has shown reliability in producing scores highly associated with human ratings (e.g., Attali & Burstein, 2006; Shermis & Hamner, 2012), the limitation in technology is a concern because formative feedback is most useful to students when it directs students' and teachers' attention to specific ways that students' writing could be improved, and also, when such feedback supports growth in skills that are especially challenging for students to master (e.g., thinking and reasoning shown in writing).

In addition to focusing primarily on holistic or surface-features, a second limitation of AWE systems is that they have traditionally focused on writing in response to open-ended prompts disconnected from a source text (Attali & Burstein, 2006; Crossley, Varner, Roscoe, & McNamara, 2013; Lee, Gentile, & Kantor, 2008; Page, 2003). An exception is *Summary Street*, which showed positive effects on the quality of students' summaries and comprehension of source texts (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Wade-Stein & Kintsch, 2004). While summarizing is an important writing skill to master (Graham & Perin, 2007), to meet new standards for writing, students should receive feedback that improves their ability to use source texts strategically in support of an argument. To this end, more recent research has found that a combination of automated systems – iSTART, an intelligent tutoring system to support reading comprehension and the Writing Pal AWE system – improved the quality of students' text-based analytic writing (Weston-Sementelli, Allen, & McNamara, 2018). Notably, this study was conducted with undergraduate students, not with young writers who by definition have different literacy learning needs.

A third explanation for why students' writing may not improve in response to automated feedback is that students may not possess the skills to successfully revise. Revising is a highly complex process (Flower, Hayes, Carey, Shriver, & Stratman, 1986). Less skilled writers in specific tend to make fewer revisions and/or revisions that do not increase the quality of their responses (Beach, 1979; Faigley & Witte, 1981; Graham, 1997; Matsumura et al., 2002; Patthey-Chavez et al., 2004). There is reason to believe, therefore, that many students would struggle to successfully implement the automated feedback they receive absent instructional intervention.

Understanding the limitations of automated feedback as a sole source of writing and revision support, researchers have begun to investigate ways to integrate AWE systems with instructional scaffolds that include games and tutorials (e.g., *Writing Pal*; Allen, Jacovina, & McNamara, 2016), as well as external support such as classroom-based interventions (Palermo & Thomson, 2018) and

teacher feedback alongside automated feedback (e.g., Wilson & Czik, 2016). These studies have shown somewhat mixed results. Palermo and Thomson (2018), for example, investigated the effect of the *NC Write* AWE system on students' essays when combined with Self-Regulated Strategy Development (SRSD) instruction –an approach designed to develop the self-regulatory and cognitive skills necessary for proficient writing (e.g., Graham, Harris, & Mason, 2005). Results showed that students who participated in the NC Write plus SRSD condition produced longer essays that contained more argumentation elements (e.g., presence of a claim) than students in an SRSD plus *NC Write* or control condition. Taking another approach, Wilson and Czik (2016) examined effects of combined teacher and automated feedback generated in the *PEG Writing* system to teacher feedback only on the quality of students' writing. Their results showed no difference in the final quality of students' drafts between conditions. Notably, teachers reported that it took them one-third less time to provide feedback in the combined (teacher plus PEG Writing) condition, and teachers' feedback focused more on substantive writing features. These findings are important because one of the hopes for AWE systems is that they would relieve teacher burden and increase students' opportunity to receive substantive feedback on their writing (see Roscoe & McNamara, 2013). In all, however, it remains an open question how best to design AWE systems, with or without teacher support, that improve students' ability to successfully implement the feedback they receive.

A significant barrier to designing AWE systems and linked interventions that support students' revision is that very little is known about revision behavior, that is, the different ways students take up feedback to improve the content of their writing. We identified only a few studies that have examined revision behavior in AWE systems. Roscoe, Snow, and McNamara (2013), for example, examined whether high school students attempted to implement the feedback messages they received, how substantive the revision was, and degree of improvement in students' essays aligned with feedback messages (e.g., if students added an introduction or modified a thesis statement in response to feedback related to essay introductions). They found that nearly all students made some attempt to revise their essays (i.e., implemented some sort of revision); however, fewer than half of the essays showed substantive improvement. In a second study of *Writing Pal*, the investigators categorized revision behavior as either word-level (e.g., lexical diversity, precision of word choices, frequency of pronouns) or document-level changes (e.g., total number of words, number of paragraphs, amount of new information provided in a sentence) (Roscoe et al., 2015). Results showed that students tended to implement more document- than word-level revisions. In a more recent deployment of *Writing Pal* college students improved their essays, but there was substantial variability among revision attempts (Roscoe, Wilson, Johnson, & Mayra, 2017). Essays with more revisions correlated positively with increase in essay score, and substantive revisions (e.g., adding new content) were associated with improvement. Finally, Zhu, Liu, and Lee (2020) investigated how middle and high school students responded to automated feedback generated by the c-rater-ML engine (Heilman & Madnani, 2013) and how their revisions related to improvements in the context of scientific argument writing. Results suggested that students with higher first-draft scores were more likely to revise, and revisions on the whole were positively related to a positive score change. Notably, among these studies, the degree of improvement across drafts was small to moderate, and importantly, they are silent with respect to what revisions 'looked like' for particular feedback messages and how those changes affected the quality of the content of students' essays.

In sum, research and development of AWE systems is still in an emergent phase with respect to assessing how the content of students' essays evolve across drafts in response to automated feedback messages, and this is notably the case for younger writers. Moreover, research to date on the effectiveness of these systems have tended to focus on overall improvement, rather than improvement aligned to particular feedback messages. Research is needed to better understand students' response to, and application of substantive feedback to guide future efforts to develop AWE systems that provide targeted revision support to students.

1.2. Present study

In the present study, we describe results of a pilot study of an AWE system, *eRevise*, designed to improve 5th and 6th grade students' use of text evidence in an argument essay. Specifically, we examine change in the overall quality of students' use of text evidence in their essays aligned with the feedback messages they received. We then examine how students implemented feedback messages in their revisions, and how these differed for essays that showed varying degrees of improvement. Because difficulty understanding the feedback could potentially impact students' revision efforts (e.g., Roscoe et al., 2017), we also investigate students' perceptions of the feedback in *eRevise*. Our goal is to provide insight into some of the specific difficulties students have in making use of feedback that could serve as useful leverage points for developing both AWE systems and writing interventions that support students' ability to improve their use of text evidence.

Our specific research questions are as follow:

RQ1: To what extent did students' use of evidence improve, from first draft to revised draft, based on *eRevise*'s automated scoring – overall, and on specific features of evidence use?

RQ2: To what extent did students' use of evidence improve, from first draft to revised draft, in line with the feedback given?

RQ3: How did students implement the feedback they received?

RQ4: How did students perceive the feedback in *eRevise*?

1.3. eRevise

1.3.1. Response-to-text assessment (RTA)

eRevise was designed to score responses and provide feedback to students on the Response-to-Text Assessment (RTA). Elsewhere, we have described RTA development and administration (Correnti, Matsumura, Hamilton, & Wang, 2012; Correnti, Matsumura, Hamilton, & Wang, 2013). In brief, the RTA was developed to create a feasible means for assessing students' ability to reason about

texts in their writing and use text evidence effectively to support their claims. To make our assessment relevant within the current research context, we aligned the RTA with the Common Core State Standards (NGAC/CCSSO, 2010). Aligned with the shift in emphasis in the CCSS on having students read and write about nonfiction texts, the RTA has been developed on nonfiction readings.

The assessment used in this pilot is based on a feature article from *Time for Kids* ("A Brighter Future" by Hannah Sachs) about the Millennium Villages Project, a United Nations-supported effort to eradicate poverty in a rural village in Kenya. The teacher reads the text aloud to students as they follow along with their own copy of the article. The purpose of the teacher reading the text to students is to help students comprehend the text. To this end, the teacher poses pre-determined questions at designated points throughout the reading. Vocabulary in the text that could potentially pose comprehension problems for students (e.g., fertilizer, tattered) also are defined. Finally, students are asked to respond to the following prompt: "Based on the article, did the author provide a convincing argument that 'winning the fight against poverty is achievable in our lifetime'? Explain why or why not with 3–4 examples from the text to support your answer."

1.3.2. Automated scoring of text evidence use

The RTA rubric for human raters focuses on five features of students' responses – evidence use, analysis, organization, academic style, and mechanics. In *eRevise*, we focus exclusively on evidence use (scored on a scale from "1 = low" to "4 = high").

eRevise's automated scoring model is based on four features we developed using natural language processing (NLP) techniques (see Matsumura et al., 2015; Rahimi & Litman, 2016; Rahimi, Litman, Correnti, Matsumura, Wang, & Kisa, 2014 for details). The features are designed to reflect the detailed criteria of the original RTA rubric. They are as follows.

(1) Number of pieces of evidence (NPE): To calculate NPE, project researchers first defined a list of main topics in the source text (i.e., the *Time for Kids* article) that were then incorporated into the AES system. These topics correspond to the ways the Millennium Villages project affected the quality of life in a village (e.g., hospital conditions, access to schools, malaria, agriculture; see Rahimi et al., 2014). The AES system uses a simple window-based algorithm with fixed window-size to calculate NPE. A window within the essay contains evidence related to a topic if it uses at least two keywords from the list of words for that topic. Each topic is only counted once to avoid redundancy.

(2) Specificity (SPC): For each main topic from the source text, researchers identified a comprehensive list of associated keywords (i.e., specific text evidence/examples). For example, the topic "hospital conditions" included as keywords "water," "electricity," "hospital beds", "medicine," and "doctors" (initially, these aspects were lacking or insufficient). For each student essay, the AES system used this keyword list to identify matches – i.e., how many (and which) specific pieces of evidence the essay addressed. The system included accounts for the similarity between a word in the student's essay and a word in the topic or key-words list, so students will be credited for evidence that uses slightly different words (e.g., "power" instead of "electricity") or words with different stems.

(3) Concentration (CON): High concentration signals listing of evidence without explanation or elaboration and receives a low score. Concentration is a binary feature. To calculate this feature, the AES system counts the number of sentences that contain keyword matches. If there are fewer than three sentences, the concentration is deemed high (i.e., undesirable).

(4) Word count (WOC): This feature is a proxy for elaboration of thinking and for students using their own language to reason how the evidence supports their main idea versus just letting the evidence speak for itself.

eRevise was developed and evaluated using 1569 Grade 4–6 essays from the RTA dataset introduced in Correnti et al. (2013). The distribution of RTA evidence scores in this corpus (assessed on a 4-point scale) was: 1 = 469 essays, 2 = 594 essays, 3 = 335 essays, and 4 = 171 essays. To evaluate the reliability of our learned scoring model, we compared the automated and human scores for our corpus using a 10-fold cross validation paradigm. That is, our sample of 1569 essays was first divided into 10 subsets, each of which was used as a test set after combining and training on the other 9 subsets. The average ICC across all 10 test sets of human- and machine-scored essays was 0.62. This level of performance approaches the human-human ICC of 0.67 for the double-scored portion of this corpus.

Recently, we have presented validity evidence supporting automated scoring of the RTA (Correnti et al., 2020). We found close correspondence between human and AES scores; at the classroom level, single measures ICC was .899, which is in the good, almost excellent, range. We also found alignment of AES scores with components of instruction that we expect would predict variation in students' writing quality. Specifically, the AES scores are sensitive to differences between classrooms in opportunities to learn analytic text-based writing, after adjusting for prior achievement, student background, and opportunities related to general reading instruction. Finally, we found associations between AES scores and other expected measures of student achievement (e.g., state standardized achievement measures in reading and math). These findings provide encouraging evidence that AES technologies as applied to the RTA can generate valid inferences about students' ability to marshal text evidence in writing. We have yet to examine construct validity at the feature level (e.g., whether features scores capture the constructs of "specificity" or "elaboration"). We note, however, that our NLP features were based on solid text-based writing constructs as articulated in the original RTA rubric. For examples of student responses that received certain scores and that demonstrate expected strengths and weaknesses, see Rahimi and Litman (2016) and also Figs. 4–7 in the present paper.

1.3.3. Automated feedback on text evidence use

eRevise uses NLP features generated during automatic scoring of students' initial essays to select formative feedback on evidence use to guide essay revision. There are three levels of feedback (see Table 1), and they adhere to features of effective feedback (Catrambone, 1998; Wang et al., 2018). For example, Level 1 feedback focuses on *completeness* (i.e., guides students to provide more evidence) and guides students to be more *specific* about the evidence they referenced. Level 2 feedback also prompts students to be

Table 1
Feedback Focus and Messages Corresponding to Each Feedback Level in *eRevise*.

Feedback Level	Feedback Focus (Code)	Feedback Messages
1 (Completeness & Specificity)	Use more evidence from the article (Completeness) Provide more details for each piece of evidence you use (Specificity)	• Reread the article and the writing prompt. • Choose at least three different pieces of evidence to support your argument • Consider the whole article as you select your evidence • Add more specific details about each piece of evidence. For example, writing "The school fee was a problem" is not specific enough. It is better to write, "Students could not attend school because they did not have enough money to pay the school fee." • Use your own words to describe the evidence
2 (Specificity & Explanation)	Provide more details for each piece of evidence you use (Specificity) Explain the evidence (Explanation)	• Add more specific details about each piece of evidence. For example, writing "The school fee was a problem" is not specific enough. It is better to write, "Students could not attend school because they did not have enough money to pay the school fee." • Use your own words to describe the evidence • Tell your reader why you included each piece of evidence. Explain how the evidence helps to make your point.
3 (Explanation & Connection)	Explain the evidence (Explanation) Explain how the evidence connects to the main idea and elaborate \(Connection)	• Tell your reader why you included each piece of evidence. Explain how the evidence helps to make your point. • Tie the evidence not only to the point you are making within a paragraph, but to your overall argument. • Elaborate. Give a detailed and clear explanation of how the evidence supports your argument.

more *specific*, and it directs students to *explain* their evidence. Finally, Level 3 feedback focuses students on not only *explaining* the evidence they provided, but also *connecting* it to the overall argument. To determine the appropriate feedback to provide to each essay, *eRevise* draws on a feedback selection algorithm. The algorithm takes into account both the NPE value and a count of the number of non-duplicate, unique evidence words or phrases from four primary topics in the SPC feature. For example, a student whose essay has low NPE and low SPC count would be asked to provide more pieces of evidence, and ensure the evidence they add are detailed (Level 1 feedback). Meanwhile, an essay with sufficient number of pieces of evidence and high SPC count would be encouraged to revise with a focus on explanation and connection (Level 3 feedback; see Zhang et al., 2019 for technical details).

2. Methodology

2.1. Context

This pilot study took place in two public parishes (i.e., districts) in Louisiana. Students in 3rd to 8th grade take the Louisiana Educational Assessment Program (LEAP) 2025 ELA test, which is aligned to the Louisiana Student Standards, based heavily on the CCSS. The assessment features prose constructed response (PCR) writing tasks requiring students "to show understanding of text(s) by writing a multi-paragraph response" that uses evidence from the text(s). Also, among other dimensions, students' literary analysis response is scored for use of "clear reasoning supported by relevant text-based evidence in the development of the topic."

Both parishes in which the study was situated are rural. In one parish, about 73 % of the student population is White; 20 % is Black, and the remaining 7 % identify as Hispanic, Asian, American Indian, Hawaiian/Pacific Islander, or multiple races. In the other parish, about 69 % of the student population is White; 15 % is Black, 8 % Hispanic, and the remaining 8 % identify as Asian, American Indian, Hawaiian/Pacific Islander, or multiple races. About 65 % and 55 % qualify for free or reduced-price lunch. About 1 % of students in each parish are English Language Learners; about 9 %–12 % have IEPs (i.e., qualify for special education services).

2.2. Participants

2.2.1. Teachers

Seven ELA teachers from seven different schools participated in the study in the 2017–2018 school year. They were selected for their comfort level with basic use of technology and access to a class set of computers for student use to complete the RTA. All seven teachers were white females holding a Bachelor's degree; one also held a Master degree in Education. They averaged 10 years (range = 4–18) of teaching experience. Three teachers taught 5th grade ELA; three taught 6th grade ELA, and one taught both 5th and 6th-grade ELA.

2.2.2. Students

The seven teachers administered the RTA via *eRevise* to all students in one of their ELA classes (n = 160). The classes averaged 23 students (range = 7–36). In the end, 143 students completed all data collection (i.e., submitted both a first draft and a revised draft of the essay).

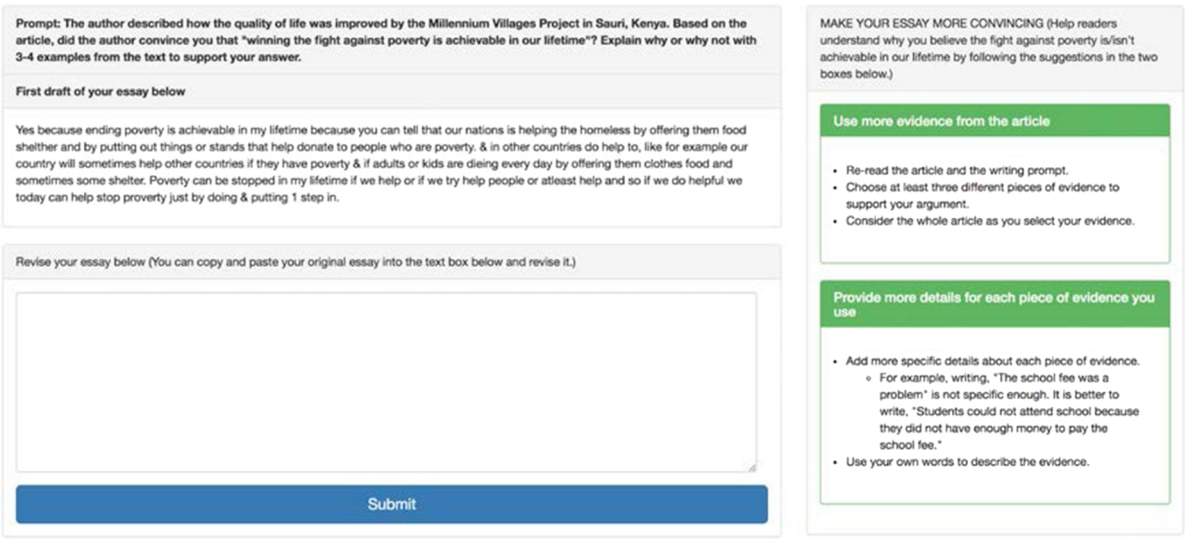

Fig. 1. *eRevise* screenshot. The screen is split into three parts. The left top box shows a student's first draft. This helps students to recall their first draft and eases revising or re-writing (e.g., by allowing cutting and pasting). The right-hand side of the screen shows the feedback on the first draft that was automatically selected by the system. The left bottom box shows where students create their second drafts, hopefully guided by the feedback displayed on the right.

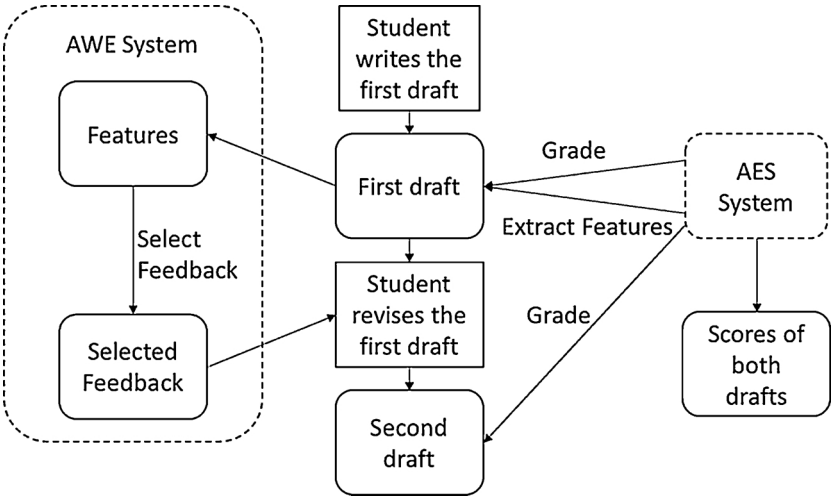

Fig. 2. Architecture of *eRevise*.

2.3. Procedures

Teachers implemented *eRevise* in May, after administration of the LEAP assessments. The *eRevise* system is designed for use over two class periods. Students wrote (i.e., typed) their essays on the first day. After students submitted their first drafts, *eRevise* extracted features for the feedback selection algorithm. It selected one of the three feedback levels that best addressed the problems of the draft. On a second day (no more than five school days later), students logged into *eRevise* to revise their original drafts using the formative feedback produced by *eRevise*. Fig. 1 shows an example screenshot with formative feedback that students would see on day 2. Note that while *eRevise* generates an automated score in the background, students do not receive the score at any point. See Fig. 2 for a depiction of the overall architecture of *eRevise*.

Teachers were instructed to provide at least 30 min of independent work time on day 1 for students to draft their essay, and on day 2 for them to revise. Actual revision times varied within and across classes. According to *eRevise*'s built-in time log, the average revision time across classes was approximately 25 min (range = 13−57 min)[1] . Students completed a brief survey upon submitting

[1] The elapsed time is a rough estimate of the amount of time students spent revising their essay. We cannot be certain that students began working as soon as they logged into *eRevise,* nor that they worked without interruption from the time they logged in until the time they logged off.

their final drafts on their understanding and perception of the usefulness of *eRevise*'s feedback.

2.4. Data sources and data analysis

To address our first three research questions, we collected, coded, and analyzed both versions of the student essays (i.e., first draft and revised):

2.4.1. RQ1 analysis

Improvement in drafts based on the *eRevise* system was assessed by conducting paired-sample t-tests to compare evidence scores for the first and revised drafts. Specifically, we compared the automated scores for each piece as well as specific feature scores (i.e., NPE, SPC).

2.4.2. RQ2 analysis

We conducted qualitative analysis of the revisions students made between drafts to characterize the extent to which students' revisions reflected the feedback they received via *eRevise*. To do this, we first created a Word file with all first drafts and a file with all second drafts, ordered by a unique student ID. We then used the 'Compare Documents' function to create a file that tracked changes between the two versions (i.e., shows the insertions, deletions, and other revisions that students made).

With knowledge of the feedback each essay received (i.e., Level 1, 2, or 3), we first coded for *whether the students attempted to revise their essay* based on the feedback provided. In other words, did students' revisions reflect at least one of the feedback topics or focus[2] (see Table 1)? For essays receiving Level 1 feedback focused on completeness and specificity ("Use more evidence from the article" and "Provide more details for each piece of evidence you use"), for example, we looked for revisions where students attempted to add a piece of evidence that was not previously mentioned and/or added more specificity to a piece of existing evidence[3] . For Level 3 feedback focused on explanation and connection to the overall argument ("Explain the evidence" and "Explain how the evidence connects to the main idea and elaborate"), we looked for signals of attempts to provide an explanation for pieces of evidence (e.g., phrases like "I included this because…" or "This evidence means…") and/or connection between the evidence and the overall argument (i.e., why students believe poverty can or cannot end in their lifetime).

Then, for the essays for which students attempted revisions, we coded the *extent to which the essays showed improvement* from first to second draft, considering the feedback provided. This meant we evaluated essays that were prompted to provide more pieces of evidence on the amount and quality of additional evidence, not on other types of revisions (e.g., extended explanations of evidence) that may have improved the essay. We did this because our primary interest was the extent to which the feedback *eRevise* provided guided students to execute successful revisions (i.e., the extent to which students revised according to the feedback they received).

We coded essays as showing *no improvement* in evidence use if apparent attempts to implement feedback resulted in no or very superficial revisions. For example, students may have repeated a piece of evidence they already provided, or used the phrase, "This evidence means…" to signal an attempt to provide explanation, but what followed was not, in fact, an explanation. Essays showed *slight improvement* if revisions were in the direction of the feedback provided, but were inconsistently implemented or not fully executed. For instance, students may have added one, instead of multiple, pieces of evidence, or added one brief phrase to explicate a piece of evidence instead of fully articulating how the evidence supports the point. Finally, we deemed essays to show *substantive improvement* if the revisions markedly improved the use of evidence (i.e., if the evidence was significantly more complete, accurate, specific, and/or explained) (Wang, Matsumura, & Correnti, 2017, 2018). Note though, that successfully revised essays may not necessarily exemplify evidence use; they could still be further improved.

Overall then, to address RQ2, four codes were possible: No attempt at implementing given feedback; Attempted to implement feedback, but no improvement in evidence use; Slight improvement in evidence use; and Substantive improvement in evidence use (see Table 4 in Findings section).[4] The first author coded all 143 pieces of student work; the second author double-coded 35 pieces (24 %). A 86 % exact agreement and 91 % adjacent agreement was reached; Cohen's kappa was 0.77, indicating 'substantial' agreement (McHugh, 2012). We discussed discrepancies to arrive at consensus and to establish or clarify decision rules to guide future coding.

2.4.3. RQ3 analysis

To diagnose how students implemented the feedback they received, or in other words, why students' attempt to revise their essay was effective or ineffective, we performed qualitative, iterative, and inductive coding. For those essays that showed *no attempt* at implementing feedback, we derived three codes that represented the different ways essays appeared not to have attempted revisions (see Table 5). For essays that resulted in *no or slight improvement* in evidence use, we coded how students' revisions missed the mark,

[2] We allowed that, given two feedback topics or focus and time constraints, students may choose to or only manage to focus on one and not both in the revision process.

[3] Recall that we had an expert-generated list of evidence topics and key words from the source text that we anticipated students would marshal in their essay.

[4] Our coding appears to differ from Roscoe et al. (2013). They coded first for whether students attempted to revise by making any edits, not necessarily edits aligned to feedback provided. Then, they coded whether students attempted substantive revisions. Any attempt to compare our results with those of Roscoe et al. (2013) should be done with caution.

Table 2
Distribution of First Draft *eRevise* Essay Scores and Feedback Level Provided (n = 143).

First-Draft *eRevise* Score	Feedback Level 1 (n = 45)		Feedback Level 2 (n = 27)		Feedback Level 3 (n = 71)		Total	
	N	%	N	%	N	%	N	%
1 (low)	24	17 %	6	4 %	1	1 %	31	22 %
2	13	9 %	20	14 %	9	6 %	42	29 %
3	4	3 %	1	1 %	17	12 %	22	15 %
4 (high)	4	3 %	0	0 %	44	31 %	48	34 %

as in fell short of actualizing substantive improvements in the revised draft. We arrived at a total of 12 codes in two broad categories representing the two primary feedback foci and the two broad categories of revisions students attempted: Providing more evidence or more detailed evidence, and explaining or connecting the evidence (see Table 6). The specific codes represented the reasons we assessed the revisions as unsubstantive. Finally, we coded essays wherein revisions showed attention to the feedback provided and resulted in *substantive improvement*. Our codes represented how the revisions were successful (see Table 7). Each essay received one code; coders selected the most applicable code.

Once again, for the RQ3 analyses, the first author coded all the essays (n = 143). The second author double-coded 35 pieces (24 %), and a 77 % exact agreement was reached; Cohen's kappa was 0.76, indicating 'substantial' agreement (McHugh, 2012). We discussed discrepancies to arrive at consensus and made decision rules to guide future coding.

2.4.4. RQ4 analysis

Items on the student survey focused on their perception and use of the feedback in *eRevise* were assessed on a four-point scale (1 = Not at all/None, 4 = Completely/All). Descriptive statistics were used to characterize the range of students' responses.

3. Findings

3.1. RQ1: To what extent did students' use of evidence improve, from first draft to revised draft, based on eRevise's automated scoring?

Of the 143 pieces of first-draft essays *eRevise* scored, 31 (22 %) received a score of 1 (lowest); 42 (29 %) received a score of 2; 22 (15 %) received a score of 3; and the remaining 48 (34 %) received a score of 4 (highest). Also, of all the essays, 45 (31 %) received Level 1 feedback focused on completeness and specificity; 27 (19 %) received Level 2 feedback focused on specificity and explanation; and 71 (50 %) received Level 3 feedback focused on explanation and connection to the overall argument. Table 2 summarizes the distribution of scores.

Students showed significant improvement in their estimated evidence score on their first drafts (M = 2.61, SD = 1.16) compared to their revised draft scores (M = 2.78, SD = 1.15; t(142) = 3.31, p = .001, ES = .15). These findings were observed despite the fact that 48 out of the 143 first-draft essays (34 %) were scored a "4" (on a scale of 1 = low to 4 = high), leaving no room for improvement in our AES system. For the students whose first-draft essay scores were less than "4" (n = 95), there was obviously greater improvement in their evidence rubric score from their first drafts (M = 1.91, SD = .74) to their revised draft scores (M = 2.24, SD = .99; t(94) = 5.52, p < .001, ES = .39).

The scatterplot in Fig. 3 shows that the proportion of students who improved (20 %; n = 29) is greater than those who declined

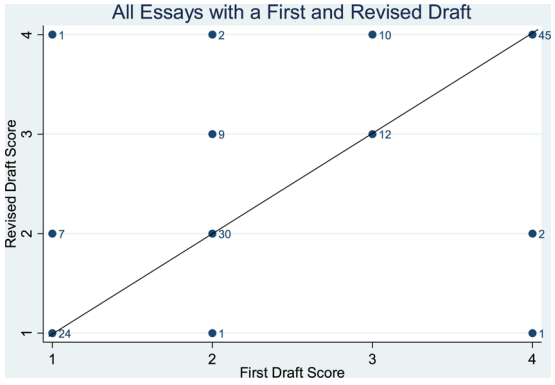

Fig. 3. Scatter plot of first draft scores by revised draft scores following automated feedback for 143 students.
Note: The diagonal line represents essays where the scores on the first and revised drafts received *the same* automated score. Above the line are essay scores demonstrating improvement on the revised draft over their first draft, and below the line are students' essays that were lower on the revised draft than they were on their original draft.

Table 3
Paired-Samples *t*-test Results for Outcomes Indicating Improvement from First- to Revised- Drafts.

Outcome	All Essays (n = 143)					Essays with Evidence Score < 4 on 1st Draft (n = 95)				
	1st Draft M (sd)	Rev. Draft M (sd)	t	p-value	ES	1st Draft M (sd)	Rev. Draft M (sd)	t	p-value	ES
Malaria-Related Text Evidence	2.26 (1.70)	2.55 (1.68)	4.19	.000	.17	1.62 (1.43)	1.98 (1.46)	4.30	.000	.25
Hospital-Related Text Evidence	2.38 (1.96)	2.78 (2.06)	5.01	.000	.20	1.82 (1.78)	2.23 (1.93)	4.60	.000	.22
Agriculture-Related Text Evidence	1.24 (1.52)	1.38 (1.51)	2.04	.044	.09	0.56 (0.92)	0.80 (1.13)	2.87	.000	.33
School-Related Text Evidence	1.96 (1.74)	2.40 (1.71)	5.11	.000	.26	1.38 (1.51)	1.94 (1.56)	4.82	.000	.36
Improving Conditions-Related Text Evidence	5.50 (4.09)	5.99 (3.98)	3.78	.000	.12	3.48 (2.94)	4.16 (2.98)	4.66	.000	.23
Cumulative Text Evidence in Focal Topics (Sum)	13.34 (8.31)	15.09 (8.20)	5.96	.000	.21	8.86 (5.44)	11.11 (5.83)	6.38	.000	.40
Breadth of Text Evidence (NPE)	2.61 (1.27)	2.81 (1.08)	3.33	.001	.17	2.14 (1.24)	2.46 (1.05)	4.09	.000	.26
Word Count	249.6 (118.3)	328.6 (165.0)	11.91	.000	.56	200.5 (90.9)	264.8 (124.1)	11.52	.000	.60

Note: t = t-statistic; ES = Effect Size.

(3%; n = 4), although the vast majority (77 %; n = 110) had the same revised and first draft scores. If we remove those 48 cases whose first draft scores were "4", then 29 of the 95 student essays (31 %) improved. These 29 student essays were distributed across different feedback levels; with 37 % receiving Level 1 feedback (on completeness and specificity) improving; 21 % receiving Level 2 feedback (on specificity and explanation) improving, and 33 % receiving Level 3 feedback (on explanation and connection) improving. Thus, some student essays at each feedback level showed improvement.

To address the question of what proportion of essays improved on specific features of evidence use, we used finer-grained outcomes. This analysis allowed us to assess potential improvements in evidence features that *eRevise* extracted (i.e., NPE, SPC), even if the revisions did not result in an overall score change, which could have been the case for essays that received a first-draft score of 4. On all tested outcomes, using all 143 essays, revised essays showed significant improvements in the number of pieces of evidence students included for five central source-text topics (see Table 3). These outcomes are elemental to the automatically scored features (e.g., NPE) used in scoring and deciding feedback levels.

Table 3 shows that students' revisions do show increases in effect size (ES) for word count (ES = .56 for all essays and ES = .60 for essays where the first draft scored less than 4 on the rubric), as well as *number of references* to specific evidence in the text, as identified through our AES scoring (ES ranging from .09 to .26 for all essays; and ES ranging from .22 to .40 for essays where the first draft scored less than 4). This analysis constitutes one way of thinking about change in evidence use from the first to revised draft – by counting the presence of specific pieces of evidence. It does not, however, evaluate the degree to which those pieces of evidence necessarily improved the quality of the essay, where quality takes into account the relevance of the evidence to the claim and reason and the explanation given. To assess this, we turn to the qualitative coding work.

3.2. RQ2: To what extent did students' use of evidence improve, from first draft to revised draft, in line with the feedback given?

Table 4 presents the results of our qualitative coding of students' revisions. Out of 143 essays, 20 % (n = 28) showed *no attempt* to implement the feedback *eRevise* provided. About a third of the essays (34 %; n = 48) exhibited some attempt to implement the feedback, but there was *no noticeable improvement* in essay quality in line with the feedback given. About another third of the essays (29 %; n = 41) showed *slight improvements*, and only about 18 % (n = 26) of the essays demonstrated *substantive improvements* in quality. This general pattern held for essays receiving Level 1 feedback on completeness and specificity (n = 45) and Level 2 feedback on specificity and explanation (n = 27). Among essays receiving Level 3 feedback on explanation and connection (n = 71), however, more essays *slightly improved* (39 %) or *substantively improved* (21 %). In fact, of the 41 essays that made slight improvements, 28 (68 %) received Level 3 feedback, and of the 26 essays that demonstrated substantive improvements, 15 (58 %) received Level 3 feedback.

These findings resonate with the key RQ1 findings above based on *eRevise* scoring. In both sets of analyses, about 20 % of essays improved substantively in evidence use, and the other 80 % essentially remained the same or declined in quality (i.e., made "no" or "slight improvement"). While the analysis based on *eRevise* scores suggests that more essays receiving Level 1 feedback (on completeness and specificity) improved, the qualitative analysis suggests that students whose essays received Level 3 feedback (on explanation and connection) executed more substantive and successful revisions.

3.3. RQ3: How did students implement the feedback they received?

3.3.1. Essays that did not demonstrate attempt at revising based on feedback provided

Table 5 shows that of 143 total essays, 28 did not reflect an attempt to revise the essays according to the feedback given at all. Sixteen of these essays (57 % of 28) showed revisions that smoothed out the organization, style, or language (i.e., grammar, spelling) of the essay, instead of attending to evidence use, as the feedback prompted. Nine essays (32 %) showed evidence-related revisions

Table 4
Codes, Definitions, and Results Related to the Extent to which Evidence Use Improved from First to Revised Draft, in Line with the Feedback Provided.

Code	Definition	Total (n = 143)		Of those receiving FB Level 1 (n = 45)		Of those receiving FB Level 2 (n = 27)		Of those receiving FB Level 3 (n = 71)	
		N	%	N	%	N	%	N	%
No attempt at implementing feedback	No revisions, or revisions did not reflect either of the feedback focus	28	20 %	13	29 %	7	26 %	8	11 %
Attempted to implement feedback, but no improvement in evidence use	Apparent attempt to implement feedback resulted in no or very superficial revisions	48	34 %	16	**36 %**	12	**44 %**	20	28 %
Slight improvement in evidence use	Revisions are in line with feedback provided, but inconsistently implemented or not fully executed	41	29 %	9	20 %	4	15 %	28	39 %
Substantive improvement in evidence use	Revisions markedly improved evidence use in line with the feedback provided and reflecting our framework of effective evidence use – complete, accurate, specific, and explained	26	18 %	7	16 %	4	15 %	15	21 %

Table 5
Codes, Description, and Findings Related to How Students' Essays Did Not Demonstrate Attempt at Revising Based on Feedback Provided.

Code	How essay did not demonstrate attempt at revising based on feedback provided (n = 28)	N	%
N1	Revision focused on smoothing out organization, style, or language (e.g., grammar, spelling)	16	57 %
N2	Made only revisions not keyed to the feedback received	9	32 %
N3	No change at all from draft 1 to draft 2	3	11 %

that were not aligned with the feedback given; for example, students provided more pieces of evidence when they were asked to explain how their evidence connected to the overall argument. And three essays (11 %) remained unchanged from first to second draft.

3.3.2. Essays that showed no or slight improvement

The top part of Table 6 shows that 30 of the essays that received feedback to *provide more evidence or more detailed evidence* showed no improvement or slight improvement in evidence use. In these cases, students' revisions were inclined to be unproductive in five ways. The most notable weakness was that students added very few new pieces of evidence (P1 in Table 6; 37 %). Revisions also faltered because, in trying to add more details to the evidence, students provided extraneous information – random, inappropriate, or irrelevant details – that did not serve to further the argument (P2: 33 %). In a third group of essays, the added evidence simply repeated evidence already provided in the original draft (P3: 17 %). While some revisions resulted in added details that were not text-based (P4: 7 %), others left the evidence general and vague, students added details to only some, but not all, pieces of evidence (P5: 7 %). In each case, students exhibited incomplete responses suggesting they had difficulty reflecting on their own writing, making judgments about how their writing could improve and/or applying suggestions for improvement to their writing.

In the bottom half of Table 6, we identified seven common 'missteps' among the 59 essays in which students focused their revisions on *explaining or connecting evidence to their overall argument*. First, students repeated or paraphrased the evidence instead of providing a clear explanation of how the evidence supports the claim (E1 in Table 6: 39 %). Meanwhile, students made revisions wherein they added personal commentary instead of connecting the evidence to the claim or larger argument (E2: 19 %). Other students added a one- or two-sentence conclusion that did not fulfill the function of connecting the evidence to their argument (E3: 17 %). Whereas some students' revisions did not improve their argument because students recycled the same pro forma explanation for each piece of evidence (E4: 7 %), other students' explanations for their evidence focused on the choice of evidence as an academic exercise instead of the content of the evidence they provided (E5: 7 %). In some cases, students' revisions were inadequate because they only added explanation to some, but not all, of the evidence provided, or they wrote a blanket explanation for all pieces of evidence (E6: 7 %). Finally, a few students, in trying to add an explanation, students instead summarized the source text (E7: 5 %). In each case, students' inability to effectively use evidence by connecting it to their reasons and/or main claim led to limited improvement in their overall argument.

3.3.3. Essays that showed substantive improvement

The top part of Table 7 shows that nine students successfully revised their essays by *providing more evidence or details*. Six of these students (SP1: 67 %) added more than one relevant piece of evidence that had not been in the original draft. Two students (SP2: 22 %) added substantive details to evidence already presented, thereby improving specificity. Finally, one student successfully implemented the feedback given by not only adding more relevant text-based examples, but also by deleting examples and generalizations based on personal knowledge or assumptions rather than the source text.

The bottom half of Table 7 shows that 17 essays demonstrated substantive improvement with respect to *explaining evidence or connecting evidence to the overall argument.* Most prominently, during revision, eight students provided an explanation for each piece of evidence, whereas previously, there had been no explanation at all, or only a paraphrase of the evidence (SE1: 47 %). In five essays (SE2: 29 %), students enacted successful revisions by connecting the evidence to the overall argument, for example, through statements reasoning that the progress made in the Kenyan village over a short period of four years suggests that poverty can be ended in the students' lifetime, which spans many more decades. Finally, in four essays (SE3: 24 %), students took the revision opportunity to elaborate upon the explanation they had attempted to provide in the original draft; they fleshed out their ideas more, beyond a short phrase.

3.4. Annotated examples of first-draft and revised essays

In this section, we provide brief annotated examples of four first-draft and revised-draft responses to illustrate some of the findings described above. These examples serve two primary purposes. First, they help to lay bare our methodology and coding. In particular, our rating of the extent of improvement (i.e., no, slight, substantive) is at the essay level; with these examples, readers can appraise our assessments of the quality of students' revisions. Second, to date, there have been minimal insights about what content improvement looks like for young writers in the context of AWE systems. By featuring such examples, we begin to provide insight into the potential range of development in students' revision abilities.

First, we present an example of an essay that received Level 1 feedback (i.e., to use more evidence) that was unsuccessful in the revisions. Then, we present a counter-example of an essay that received the same feedback, but that demonstrated substantive

Table 6
Codes, Description, and Findings Related to How Students' Revisions Showed No or Little Improvement in Evidence Use from First to Second Essay Draft.

Code	Why revisions showed no or little improvement in evidence use (n = 89)	N	%	Example
Of essays that primarily attempted to *provide more evidence or details* (n = 30*)				
P1	Student added very few pieces of evidence, so still insufficient	11	37 %	To a draft that mentioned only school fees, a student only added a brief mention of malaria.
P2	Student added more details to evidence, but details were random, did not serve to further the argument	10	33 %	A student added the underlined details: "The author did convince me. Because it states that "The sub District Hospital now had medicine for their patient." Another reason is that the story said That now water is connected to the hospital so now they can have water. Also they have a generator for power so thy can now have light. That is why the story convinced me that the project can work."
P3	Student added evidence, but it is just a repeat of evidence already given, without more details	5	17 %	One student added, "For an example we gave them medicen. Yes it is achevlibele because the farming stuff help make more money…". The original draft, however, already mentioned medicine and farming supplies.
P4	Student added evidence, but it is not text-based	2	7 %	One student added, "Believe it or not Malaria is actually preventable!! Unlike Cancer, Ebola, and other bad diseases from foreign countries that can't be cured (Only some types of cancer can be cured)!"
P5	Student only added details to some, but not all, pieces of evidence; some pieces of evidence were still general/vague; details were still insufficient	2	7 %	A student added the underlined details: "The Millennium Villages Project went to Sauri to help because of 20,000 kids was dying of sickens. The hospital [wasn't] that good….All the farmers corps was dying because they was so poor…."
Of essays that primarily attempted to *explain evidence or connect evidence to argument* (n = 59**)				
E1	Student added explanation that just repeated or paraphrased evidence, instead of tying it to the claim, explaining *how* it supported the claim	23	39 %	After writing, "Malaria is one disease, common in Africa, that is preventable and treatable", a student merely added, "This piece of evidence shows some of the diseases the villagers are getting that are preventable."
E2	Student added commentary, not explanation of evidence that connects to claim	11	19 %	Given the text evidence, "The people of Sauri have made amazing progress in just 8 years," a student commented, "I want to help make even more of a change. The people of Sauri deserve a better life than what they were given. We can all help them make even more progress…"
E3	Student added one- or two-sentence conclusion (and may be on the right track), but did not connect the evidence to the argument or add to the argument	10	17 %	In the final sentence of the essay, a student wrote, "So I believe that, over a course of seven years, people will be able to come out of poverty because of what this showed me. In addition, over eight years, a lot can happen to make a place better."
E4	Student recycled same explanation for each piece of evidence	4	7 %	After each piece of evidence, a student wrote, "I included [this piece of evidence] because it shows one reason how the author convinced me that 'winning the fight against poverty is achievable in our lifetime'."
E5	Student added explanation that focused on choice of evidence as an academic exercise, instead of content of evidence, with focus on making an argument	4	7 %	One student wrote, "I put this evidence because it is proof from the text that backs up the answers."
E6	Student only added explanation to some, but not all, evidence, or wrote blanket explanation for all pieces of evidence	4	7 %	After providing three pieces of evidence in succession, a student wrote, "I included each piece of evidence because each piece of evidence was caused by the Millennium Villages Project which helped win the fight against poverty. The evidence shows how their lives became better because they were able to win the fight against poverty."
E7	In attempting to connect evidence to the argument, student merely summarized the article or the evidence already provided	3	5 %	One student wrote, "In our lifetime. It only took the Millennium Village 8 years for some amazing progress. It went from no bed nets to bed nets that are used in every sleeping site."Bed nets are used in every sleeping site in Sauri."(Sachs, 9).Now the Yala Sub-District Hospital has medicine, free of charge, for all of most common diseases,but they didn't use to have any doctors, any half of the patients couldn't even afford to go. That is how the author convinced us that "winning the fight against poverty is achievable in our lifetime".

* N = 30 represents all essays that primarily attempted revisions related to providing more evidence or more detailed evidence, but only showed "no improvement" or "slight improvement" in evidence use. As Table 4 shows, of all essays receiving FB Level 1, 25 showed no or slight improvement. Additionally, of all essays receiving FB Level 2, five primarily attempted to provide more detailed evidence and resulted in no or slight improvement. Together, these are the 30 essays we analyzed here.

** N = 59 represents all essays that primarily attempted revisions related to explaining the evidence provided or connecting the evidence to the overall argument, but only showed "no improvement" or "slight improvement" in evidence use. As Table 4 shows, of all essays receiving FB Level 3, 48 showed no or slight improvement. Additionally, of all essays receiving FB Level 2, 11 primarily attempted to explain evidence and resulted in no or slight improvement. Together, these are the 59 essays we analyzed here.

Table 7
Codes, Description, and Findings Related to How Students' Revisions Showed Substantive Improvement in Evidence Use from First to Second Essay Draft.

Code	Why revisions showed substantive improvement in evidence use (n = 26)	N	%	Example
Of essays that primarily attempted to *provide more evidence or details* (n = 9*)				
SP1	Student added more than one relevant piece of evidence that was not previously featured	6	67 %	An essay previously addressed only the condition of hospitals and malaria. During revision, the student added evidence about agriculture ("On the farms the crops were dying because they didn't have any good soil") and schooling ("Sauri don't attend school because parents couldn't pay for school fees").
SP2	Student added details to evidence already presented (i.e., added specificity)	2	22 %	To the observation that the hospital "was not in good shape," for example, a student added that it "used to have no running water, electricity, and medical supplies."
SP3	Student removed non-text-based evidence (i.e., based on personal experience or knowledge)	1	11 %	The student deleted, "The first way fighting against poverty can be achieved is if you have a good mindset. If you have a good mindset then you won't have to stress about anything."
Of essays that primarily attempted to *explain evidence or connect evidence to argument* (n = 17**)				
SE1	Student provided an explanation for each piece of evidence	8	47 %	For example, one student added, "I included this piece of evidence because it shows that the hospitals in Sauri had started to install some objects that could help them fight any disease or sickness").
SE2	Student connected evidence presented to the overall argument	5	29 %	Students reasoned that the progress made in Sauri was over a short period of four years. This therefore suggests that poverty can be ended in the students' lifetime, which spans many more decades.
SE3	Student elaborated on the explanation they had already attempted to provide (e.g., fleshing it out beyond a short phrase)	4	24 %	A student added the following to his/her reasoning: "How long can we continue to give them money? Eventually we won't be able to. And even if we give them money to keep generators and water running, there are still other things that may cause problems… And when you are juggling all of the problems I listed above, it will be difficult to move on to another country and help them and expect the place you left alone to stay in better conditions."

* N = 9 represents all essays that primarily attempted revisions related to providing more evidence or more detailed evidence, and showed "substantive improvement" in evidence use. As Table 4 shows, of all essays receiving FB Level 1, seven showed substantive improvement. Additionally, of all essays receiving FB Level 2, two primarily attempted to provide more detailed evidence and resulted in no or slight improvement. Together, these are the nine essays we analyzed here.

** N = 17 represents all essays that primarily attempted revisions related to explaining the evidence provided or connecting the evidence to the overall argument. As Table 4 shows, of all essays receiving FB Level 3, 15 showed substantive improvement. Additionally, of all essays receiving FB Level 2, two primarily attempted to explain evidence and resulted in substantive improvement. Together, these are the 17 essays we analyzed here.

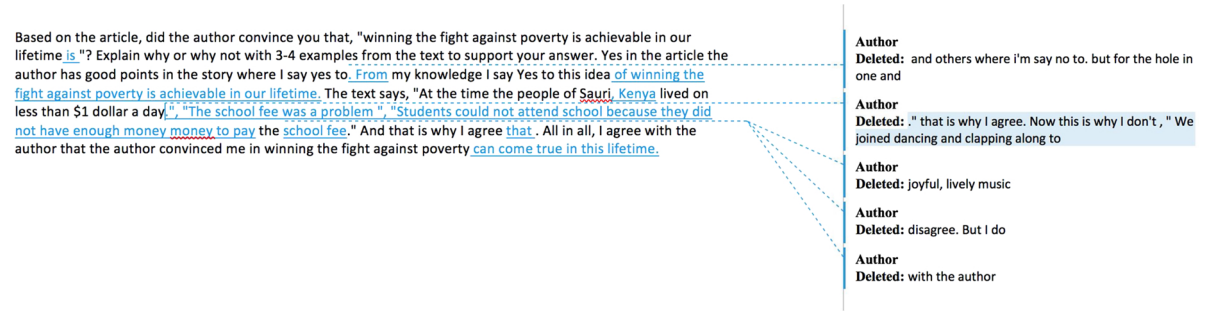

Fig. 4. Sample student essay that made no improvement in using more text evidence (Level 1 Feedback). The essay received an AES score of 1 on evidence use before and after *eRevise.*

improvement. Following this, we contrast a slightly-improved and a substantively improved essay that received Level 3 feedback (i.e., to explain evidence).

3.4.1. Essay revised according to Level 1 feedback (completeness and specificity) with "no improvement"

As shown in Fig. 4, in the original draft of this essay, the student referenced one piece of evidence from the text ("…people of Sauri, Kenya lived on less than $1 a day) in support of the author's claim that "winning the fight against poverty is achievable in our lifetime." The student also tried to provide one piece of evidence for disagreeing with the author, although the reference was actually irrelevant. During revision, the student deleted this weak piece of counter-evidence, which rendered the essay slightly clearer. Overall, however, the student did not succeed at revising according to the feedback given, which was to add more pieces of evidence. Out of the entire article with multiple other examples of poverty in the Kenyan village and how the quality of life subsequently improved (e.g., with respect to hospital conditions, malaria, agriculture), the student only referenced one more topic – the school fee. Moreover, that particular example – and in fact, the exact phrase – was given in the feedback students received (see Table 1). For these reasons, this essay was coded as demonstrating "no improvement" and the primary lapse in revision was that the student added very few pieces of evidence; text evidence use was still insufficient (P1 of Table 6).

3.4.2. Essay revised according to Level 1 feedback (completeness and specificity) with "substantive improvement"

In contrast to the first example essay, Fig. 5 shows an essay that also received Level 1 feedback to add more evidence and specific details, but that demonstrated substantive improvement upon revision. In the original draft, the student provided an appropriate and strong instance of evidence use in referencing the improvements in hospital conditions – from lacking doctors, running water, and electricity, to having medicine, water, and a generator to power the hospital. This evidence was strong because the student characterized the conditions both before and after the UN support to make the argument that Sauri has improved in just eight years, so there is hope that poverty can be eradicated in our lifetime; however, the student could have marshaled several other examples in the article to make the argument more convincing. Moreover, the intended second piece of evidence, referencing "supplies" was vague. In the revised draft, the student sharpened this general reference, clarifying that it related to school supplies, "like books, paper and pencils." In addition, the student inserted a new piece of evidence about school fees, lunches, and attendance rates that was not in the first draft. In all, the revised essay reflected understanding and execution of the feedback which further strengthened the overall argument.

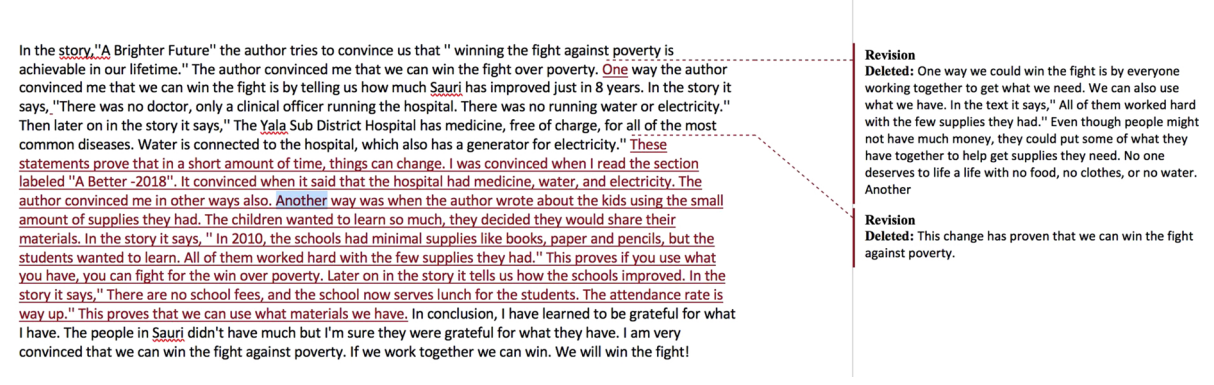

Fig. 5. Sample student essay that made substantive improvement in using more text evidence (Level 1 Feedback). The essay received an AES score of 2 on evidence use before *eRevise* and a score of 4 after.

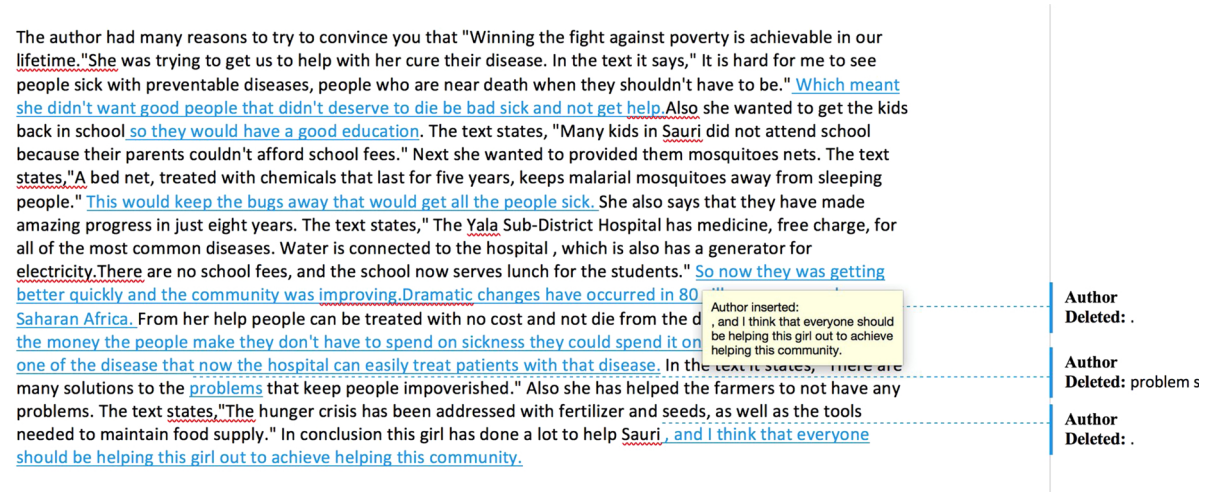

Fig. 6. Sample student essay that made slight improvement in explaining evidence provided (Level 3 Feedback). The essay received an AES score of 4 on evidence use before and after *eRevise*.

3.4.3. Essay revised according to Level 3 feedback (explanation and connection) with "slight improvement"

On the surface, the writer of the essay in Fig. 6 appeared to have significantly revised the essay by adding explanatory statements after each piece of evidence. Upon closer examination, however, we see that the student merely repeated or paraphrased the evidence, instead of explaining how it supported the claim (E1 of Table 6). For example, after providing evidence from the source article saying that "bed nets keep malarial mosquitos away from sleeping people," the student added a sentence that essentially made the same point: "This would keep the bugs away that would get all the people sick." Similarly, after highlighting the dramatic changes that occurred, the student summarized, "So now they was getting better quickly and the community was improving." A more substantive revision would have linked the evidence of the amazing progress in this one village over such a short time to the overall argument that poverty can be conquered.

3.4.4. Essay revised according to Level 3 feedback (explanation and connection) with "substantive improvement"

The essay in Fig. 7 also received feedback from *eRevise* to explain evidence and connect the evidence to the overall argument. It made some substantive revisions that resulted in an improved essay. In the first draft, the student provided four specific pieces of text

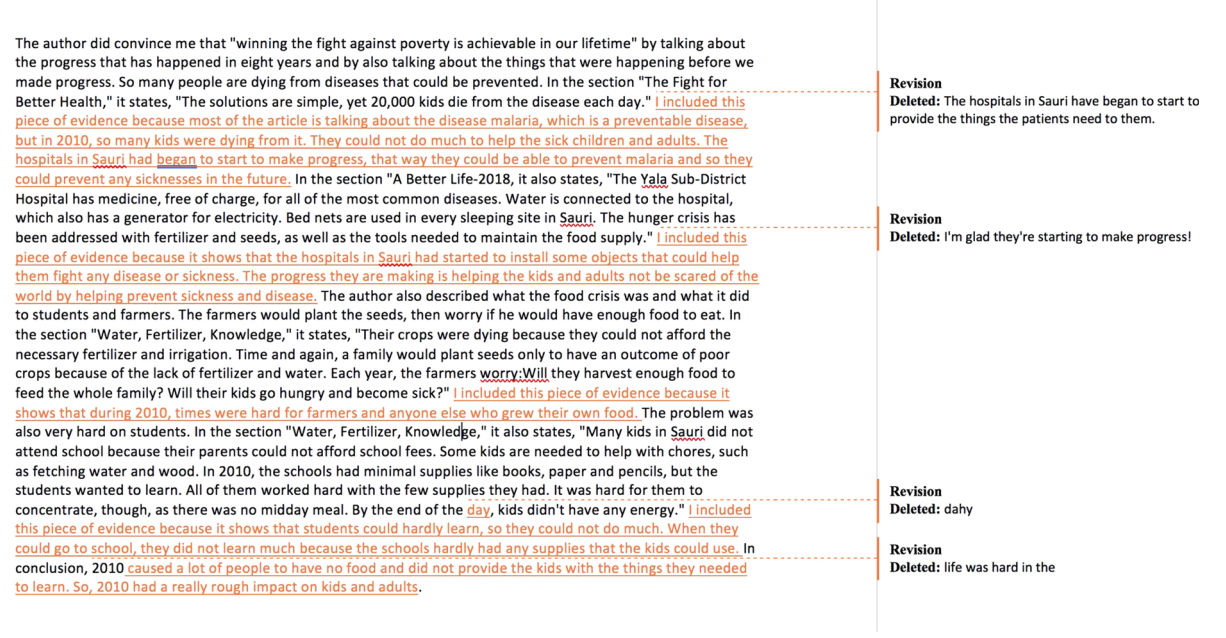

Fig. 7. Sample student essay that made substantive improvement in explaining evidence provided (Level 3 Feedback). The essay received an AES score of 4 on evidence use before and after *eRevise*.

Table 8
Student Responses to Survey Questions about their Percerption of *eRevise* Feedback (n = 136).

Question	Mean	SD	Percent of students responding			
			Not at all/ None	A little bit	Mostly/ A Lot	Completely/ All
Did you understand the feedback you received?	3.03	0.79	4 %	18 %	49 %	29 %
Did you understand how you were supposed to revise your essay based on the feedback you received?	3.08	0.79	3 %	18 %	46 %	32 %
How much of the feedback did you use when you revised your essay?	2.99	0.83	4 %	21 %	46 %	29 %

evidence, referencing all the key topics: malaria, hospital, crops, and schooling; however, these pieces of evidence were provided in succession and read like a summary of the source text. The student neglected to explain why he/she included each piece of evidence, and how it relates to the overall claim. During revision, the student added the missing explanation after each piece of evidence, albeit using a formulaic sentence starter ("I included this piece of evidence because it shows…") that echoed the Level 3 feedback's language. The additions signaled attention to the writing prompt and overall claim; it referenced the passage of time, mentioning hardships in "2010" and then progress Sauri had made to prevent sicknesses and diseases "in the future", thereby implicitly making the argument that poverty is "achievable in our lifetime." Admittedly, the connection between the evidence and the overall claim could be strengthened. In all, while the second-draft essay would still not receive the highest score for evidence use, the direction and extent of revisions renders this one of the more successfully revised essays.

3.5. RQ4: How did students perceive the feedback in eRevise?

Of the 143 students for whom we have first draft and revised essays, 136 responded to the survey questions. As summarized in Table 8, almost 80 % of students indicated that they 'mostly' or 'completely' understood the feedback they received and how they were expected to revise their feedback based on the feedback. About 75 % of the students self-reported using 'a lot' or 'all' of the feedback they received when they revised their essay. These results suggest that students thought the feedback was helpful and actionable.

4. Discussion

AES systems have long been an object of interest as a way to reduce the burden and cost of scoring summative assessments of students' writing performance. *eRevise* is attempting to explore the use of technologies for formative purposes to enhance students' knowledge of a substantive writing construct – use of text evidence – which is a critical component of academic (text-based argument) writing. Our system is unique in its focus on assessing the quality of young students' use of text evidence to support an argument, as opposed to other features of writing or general writing quality. In addition to providing insight on how to use automated essay scoring technologies to assess a substantive dimension of writing, our work also contributes to the development of essay scoring methods for younger writers whose essays are typically shorter, contain more grammatical and spelling errors, and is less sophisticated in terms of use and organization of evidence in comparison to more mature writers. Our work thus tackles the challenge of using computational techniques on data that are particularly noisy given the stage of writing development of younger students.

Our research also contributes to efforts to assess the effectiveness of AWE systems by examining the extent to which improvement in the content of students' essays is related to the feedback students received. In our study, most students' use of text evidence improved from first to second draft aligned with the feedback they received. As other AWE researchers have found, however (e.g., Roscoe, Snow, & McNamara, 2013, 2015), improvement as assessed in the *eRevise* system was small. Relatively few students (18 %) showed substantive improvement in the quality of their essays. Our results show as well that the majority of students responded positively to the system (e.g., understood what they were being asked to do), suggesting that students' unsuccessful revision attempts is likely because of a lack of skills for revising their essays – i.e., the cognitive and metacognitive skills involved with self-assessment (Nielsen, 2014) – rather than difficulty understanding the *eRevise* feedback.

We note that our study has limitations that must be considered in interpreting our results and recommendations. Chief among these is that our study was conducted in only seven classrooms. Future research needs to be conducted with larger samples of students to more robustly investigate the effectiveness of *eRevise*, as well as to ensure that we are capturing the full range of students' revision behavior. Moreover, because we did not have student-level achievement data or information about students' motivation to write, we were unable to link revision behavior to other information about students, such as whether they had reading challenges or felt particularly compelled to put effort into their writing. Finally, we note that our investigation is based on a single prompt for a single text. We are currently studying results of other tests of prompts in *eRevise*, as well as developing other methods for extracting word lists to replace manual effort (e.g., Rahimi & Litman, 2016). While our work in these directions show promise, at the present time the scalability of *eRevise* is somewhat limited.

4.1. Contributions and implications for AWE system design and research

One significant reason for the lack of research on students' implementation of feedback is that automated scoring technologies that assess the alignment between content-related feedback and revision have not yet been developed. Our study employing human raters to assess the degree of improvement between drafts with an eye toward alignment between feedback and revisions signifies a step toward that goal; that is, our codes represent categories of revision behavior that could inform future algorithms to advance argument revision analysis techniques). Automated assessment of students' implementation of feedback is critical to assessing at scale both the quality of students' writing, as well as students' ability to revise in response to feedback. Being able to do so could provide insight into potential gaps in students' understanding of evaluative criteria for good writing and so inform the design of instructional supports within AWE systems, and feedback messages that better support students' revision quality.

We believe that our findings showing the different ways that students struggled to use information from source texts in writing, also could provide information for possibly improving the interface of AWE systems. We suspect that comprehension problems might have hindered some students' revision efforts, as struggling readers often find it difficult to distinguish between relevant and irrelevant text information (i.e., can be distracted by tantalizing details that are not necessarily germane to an author's main point). For *eRevise*, we sought to circumvent potential comprehension problems by asking the teacher to read the text aloud to students before they wrote the first draft. Developers of AWE systems also might want to consider including recordings of source texts for students who might struggle with comprehension. Other supports such as highlighting examples in source texts, also could potentially be helpful for scaffolding struggling readers' ability to identify relevant information in a source text to support a claim.

Finally, an interesting pathway for future development work and research might be to investigate when in the writing process it might be most optimal to provide students with feedback. We note, for example, that our system differs from others because we provided feedback after students completed their first draft, as opposed to providing immediate feedback as students produce text, as in some other systems (e.g., Heilman & Madnani, 2013; Roscoe & McNamara, 2013). We decided to delay feedback to students because, in our view, this approach is authentic to the revising activities in which writers typically engage. In other words, whether feedback is provided by fellow students, teachers, reviewers, editors, or critical friends, writers make use of formative assessment(s) of their work – the information they receive from readers of their completed drafts – to improve the content and readability of their work. The ability to synthesize and implement comments thus is a critical skill to teach and master, and so is the focus of our AWE design. We note, however, that an argument certainly could be made for providing feedback to students while they are producing text. Moreover, because the ultimate goal of AWE systems and writing instruction that employs process approaches is to improve students' writing ability generally (i.e., not just improve one specific essay), an important area for future research would be to investigate the relative effectiveness of these approaches for supporting students' writing skills over time (i.e., their ability to transfer what is learned in AWE systems to future writing situations).

4.2. Contributions to writing instruction and research

As mentioned earlier, research on how to best teach text-based argument writing and revision skills is in an early stage. Our study provides insight into the specific ways students struggled to revise their essays, which we believe can serve as useful fodder for supporting writing instruction. As other researchers have noted (e.g., Weston-Sementelli et al., 2018), writing in response to a source text requires a high level of reading comprehension skill. Our results showing that a significant proportion of students who were prompted to add more specific evidence simply repeated the text information they had included in their first draft or, in some cases, included non-text based information suggests that many students struggled to identify pertinent information in source texts. This, in turn, suggests that an important component of teaching students how to revise – whether in an AWE system or not – might include developing norms and strategies for rereading texts with the goal of mining additional information. Such instruction would need to include, for example, helping students identify evidence they used previously in their drafts and strategically re-read the text for new examples aligned with the points they want to make.

Aligned with other writing research, our results indicated that students struggled to explain their evidence (e.g., O'Hallaron, 2014). Notably, the students who received this feedback tended to be more capable writers (i.e., their first drafts included a substantive amount of text evidence). Again, we saw a tendency for students to repeat what they had written earlier (e.g., restate their evidence) as opposed to adding explanatory text linking evidence to claims, and some students added what amounted to personal commentary disconnected from the argument. We also saw students add conclusions or other essay elements that improved their essay in certain respects, but not aligned to the feedback they received to add explanation, or apply a pro forma explanation for each piece of evidence. While students may be getting the message that they have to use evidence to support a claim, teaching students to link claims and evidence is difficult and so may be a heavier 'lift' for classroom instruction. A critical focus of writing instruction then would be making explicit to students what it means to link claims and evidence and model this for students.

Funding

This work was supported by the Institute of Education Sciences, Award #R305A160245. The opinions expressed in this article are those of the authors, not the sponsors. The authors remain responsible for any errors in the work.

Declaration of Competing Interest

We verify that this manuscript has not been published elsewhere and is not under consideration by another journal. We also confirm that there are no known conflicts of interest associated with this research.

References

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (pp. 316–329). (2nd ed.). New York: Guilford.

Attali, Y. (2004). Exploring the feedback and revision features of criterion. *Journal of Second Language Writing, 14*, 191–205.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning, and Assessment, 4*(3).

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58.

Beach, R. (1979). The effects of between-draft teacher evaluation versus student self-evaluation on high school student's revising of rough drafts. *Research in the Teaching of English, 13*(2), 111–119.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education Principles Policy and Practice, 5*(1), 7–74.

Brindle, M., Graham, S., Harris, K. R., & Hebert, M. (2016). Third and fourth grade teacher's classroom practices in writing: A national survey. *Reading and Writing, 29*(5), 929–954.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology General, 127*(4), 355.

Correnti, R., Matsumura, L. C., Hamilton, L. S., & Wang, E. (2012). Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment, 17*(2–3), 132–161.

Correnti, R., Matsumura, L. C., Hamilton, L. S., & Wang, E. (2013). Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal, 114*(2), 142–177.

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., & Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*.

Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. July *International Conference on Artificial Intelligence in Education* (pp. 269–278).

Deane, P. (2013). On the relation between automated essay scoring and modern views of thewriting construct. *Assessing Writing, 18*, 7–24.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414.

Flower, L. S., Hayes, J. R., Carey, L., Shriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication, 37*, 16–55.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33*(1), 53–80.

Graham, S. (1997). Executive control in the revising of students with learning and writing difficulties. *Journal of Educational Psychology, 89*(2), 223.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools: A report to Carnegie Corporation of New York*Washington, DC: Alliance for Excellent Education.

Graham, S., & Sandmel, K. (2011). The process writing approach: A meta-analysis. *The Journal of Educational Research, 104*(6), 396–407.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing, 27*(6), 1015–1042.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*(2), 207–241.

Graham, S., Harris, K. R., & Santangelo, T. (2015). Research-based writing practices and the Common Core: Meta-analysis and meta-synthesis. *Elementary School Journal, 115*(4), 498–522.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547.

Heilman, M., & Madnani, N. (2013). *ETS: Domain adaptation and stacking for short answer scoring. Second Joint Conference on Lexical and Computational Semantics (*SEM). Proceedings of the Seventh International Workshop on Semantic Evaluation, Vol. 2*, 275–279. Retrieved from http://www.aclweb.org/anthology/S13-2046.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Thousand Oaks, CA: Corwin Press.

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*(2), 173–196.

Lee, Y. W., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL® CBT essays: Scores from humans and e-rater®. *ETS Research Report Series, 2008*(1), i–71.

Matsumura, L. C., Correnti, R., & Wang, E. (2015). Classroom writing tasks and students' analytic text-based writing. *Reading Research Quarterly, 50*(4), 417–438.

Matsumura, L. C., Patthey-Chavez, G. G., Valdés, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*(1), 3–25.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica: Biochemia medica, 22*(3), 276–282.

National Center for Education Statistics (2012). *The nation's report card: Writing 2011 (NCES 2012–470).* NCES.

National Commission on Writing for America's Families, Schools, and Colleges (2004). *Writing: A ticket to work. . . Or a ticket out: A survey of business leaders.* September, Retrieved fromNew York, NY: College Entrance Examination Board. http://wwwwritingcommission.org/prod_downloads/writingcom/writing-ticket-to-work.pdf.

National Governors Association Center for Best Practices & Council of Chief State School Officers (NGAC/CCSSO) (2010). *Common core state standards English language arts standards.* Washington, DC: Author.

Nielsen, K. (2014). Self-assessment methods in writing instruction: A conceptual framework, successful practices and essential strategies. *Journal of Research in Reading, 37*(1), 1–16.

O'Hallaron, C. L. (2014). Supporting fifth-grade ELLs' argumentative writing development. *Written Communication, 31*(3), 304–331.

Olson, M. W., & Raffeld, P. (1987). The effects of written comments on the quality of student compositions and the learning of content. *Reading Psychology: An International Quarterly, 8*(4), 273–293.

Page, E. B. (2003). *Project essay grade: PEG. Automated essay scoring: A cross-disciplinary perspective*.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255–270.

Patthey-Chavez, G. G., Matsumura, L. C., & Valdes, R. (2004). Investigating the process approach to writing instruction in urban middle schools. *Journal of Adolescent & Adult Literacy, 47*(6), 462–476.

De La Paz, S., Ferretti, R., Wissinger, D., Yee, L., & MacArthur, C. (2012). Adolescents' disciplinary use of evidence, argumentative strategies, and organizational structure in writing about historical controversies. *Written Communication, 29*(4), 412–454.

Rahimi, Z., & Litman, D. J. (2016). Automatically extracting topical components for a response-to-text writing assessment. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications,* 277–282.

Rahimi, Z., Litman, D. J., Correnti, R., Matsumura, L. C., Wang, E., & Kisa, Z. (2014). Automatic scoring of an analytical response-to-text assessment. June *International Conference on Intelligent Tutoring Systems* (pp. 601–610).

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing stra-tegy tutor in the high school classroom. *Journal of Educational*

Psychology, 105(4) 1010.
Roscoe, R. D., Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). Automated detection of essay revision patterns: Applications for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning, 10*, 59–79.
Roscoe, R. D., Snow, E. L., & McNamara, D. S. (2013). Feedback and revising in an intelligent tutoring system for writing strategies. July *International Conference on Artificial Intelligence in Education* (pp. 259–268).
Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70*, 207–221.
Shanahan, T. (2015). Common core state standards. *Elementary School Journal, 115*(4), 464–479.
Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. April *Annual National Council on Measurement in Education Meeting,* 14–16.
Shermis, M. D., Garvan, C. W., & Diao, Y. (2008). *The impact of automated essay scoring on writing outcomes.* Online Submission.
Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.
Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333–362.
Wang, E., Matsumura, L. C., & Correnti, R. (2017). Written feedback to support students' higher level thinking about texts in writing. *The Reading Teacher, 71*(1), 101–107.
Wang, E., Matsumura, L. C., & Correnti, R. (2018). Student writing accepted as high-quality responses to analytic text-based writing tasks. *The Elementary School Journal, 118*(3), 357–383.
Weston-Sementelli, J. L., Allen, L. K., & McNamara, D. S. (2018). Comprehension and writing strategy training improves performance on content-specific source-based writing tasks. *International Journal of Artificial Intelligence in Education, 28*(1), 106–137.
Wilson, J., & Andrada, G. N. (2016). *Using automated feedback to improve writing quality: Opportunities and challenges. Handbook of research on technology tools for real-world skill development.* IGI Global.
Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109.
Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L. C., ... Quintana, R. (2019). *eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing.* July*Proceedings of the AAAI Conference on Artificial Intelligence*, *Vol. 33*, 9619–9625.
Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education,* 143.

Elaine Lin Wangis a Policy Researcher at the RAND Corporation. She specializes in using qualitative methods to examine the quality of literacy – particularly writing – instruction and learning, and to understand factors that facilitate or pose challenges for implementation of education policies, programs, or technologies.

Lindsay Clare Matsumura is a Professor in the University of Pittsburgh's School of Education (SOE) and Senior Scientist at the Learning Research and Development Center (LRDC). She studies professional development interventions for teachers, and the relationship between classroom writing tasks or discourse and literacy learning.

Richard Correnti is an Associate Professor in the University of Pittsburgh's SOE and Research Scientist at the LRDC. He studies how policy and educational reform initiatives can improve instruction and student learning, and how these efforts are influenced by issues of implementation and scaling-up.

Diane Litman is a Professor of Computer Science at the University of Pittsburgh and a Senior Scientist at the LRDC. Her research is in areas including artificial intelligence, computational linguistics, knowledge representation and reasoning. Her most recent research has been in speech and natural language technology for educational applications.

Haoran Zhang is graduate student in the University of Pittsburgh's Computer Science Ph.D. program. His main interests are machine learning and natural language processing.

Emily Howe is graduate student in the University of Pittsburgh's Learning Sciences and Policy Ph.D. program with interests in improving learning outcomes in reading, writing, and critical thinking.

Ahmed Magooda is graduate student in the University of Pittsburgh's Computer Science Ph.D. program. His main interests are machine learning and natural language processing.

Rafael Quintana is graduate student in the University of Pittsburgh's Learning Sciences and Policy Ph.D. program with interests in improving learning outcomes in reading, writing, and critical thinking.